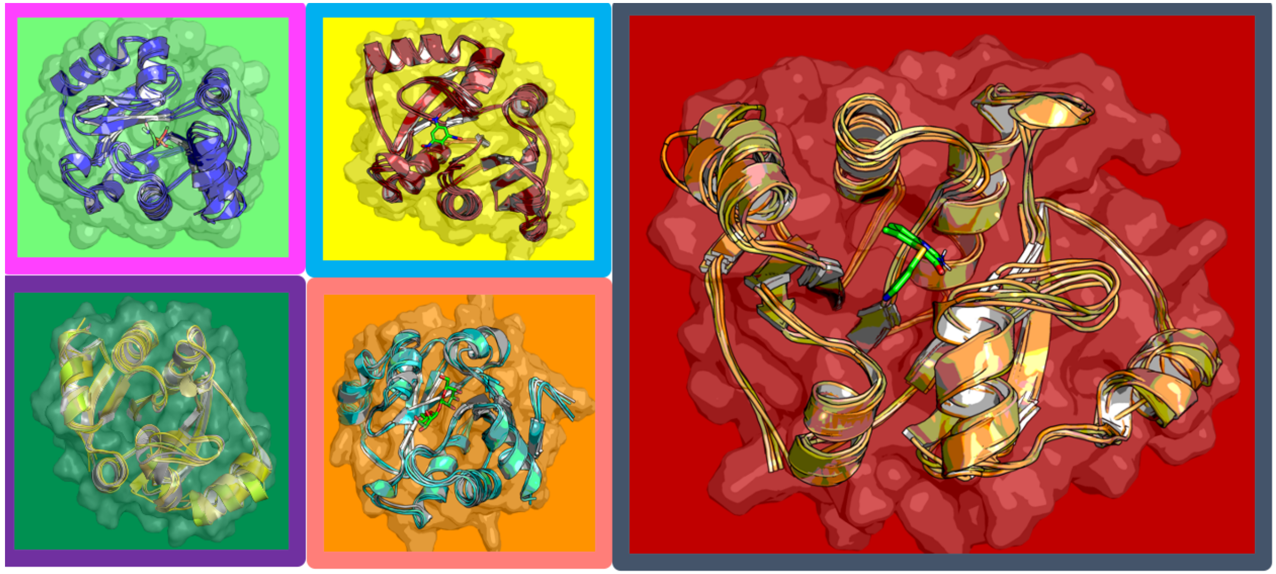
TARGETING COVID-19 PROTEOME WITH AI & MULTISCALE SIMULATIONS



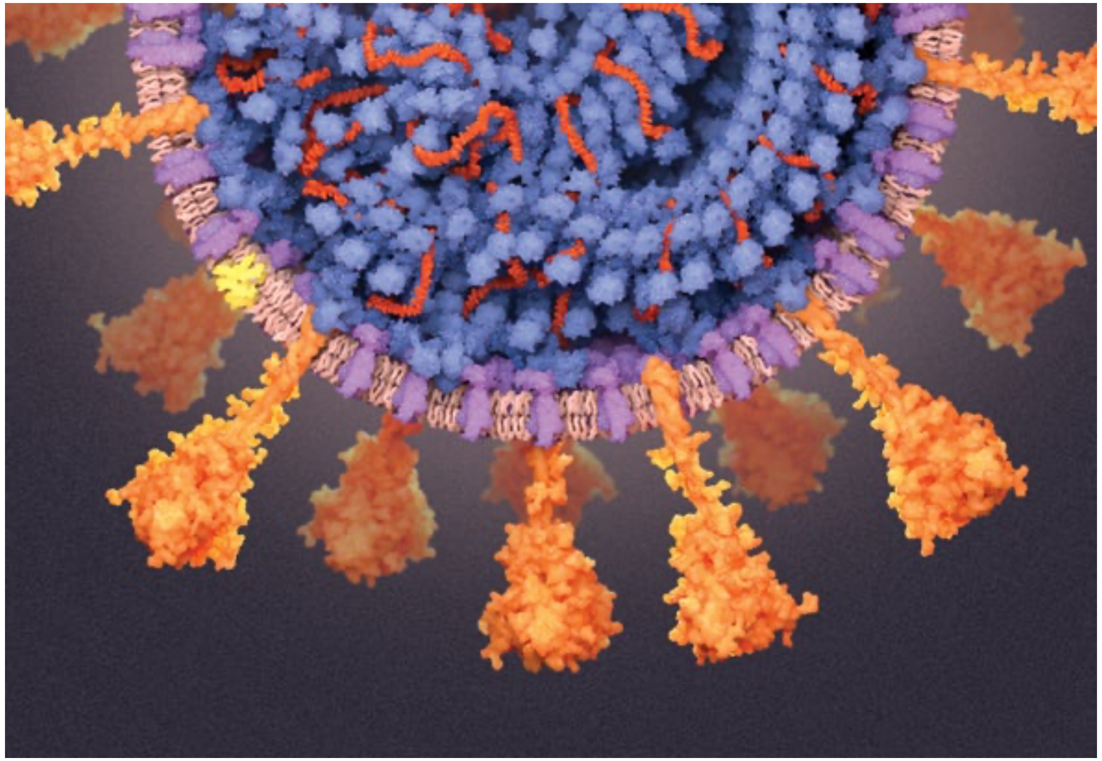
ARVIND RAMANATHAN (ON BEHALF OF THE COVID-19 TEAM – ANL + BNL)
Data Science & Learning Division, Computing, Environment and Life Sciences, Argonne National Laboratory, Lemont, IL 60439
CASE, University of Chicago
http://ramanathanlab.org
ramanathana@anl.gov

U.S. DEPARTMENT OF ENERGY
Argonne National Laboratory is a
U.S. Department of Energy laboratory
managed by UChicago Argonne, LLC.

Mar 18, 2020

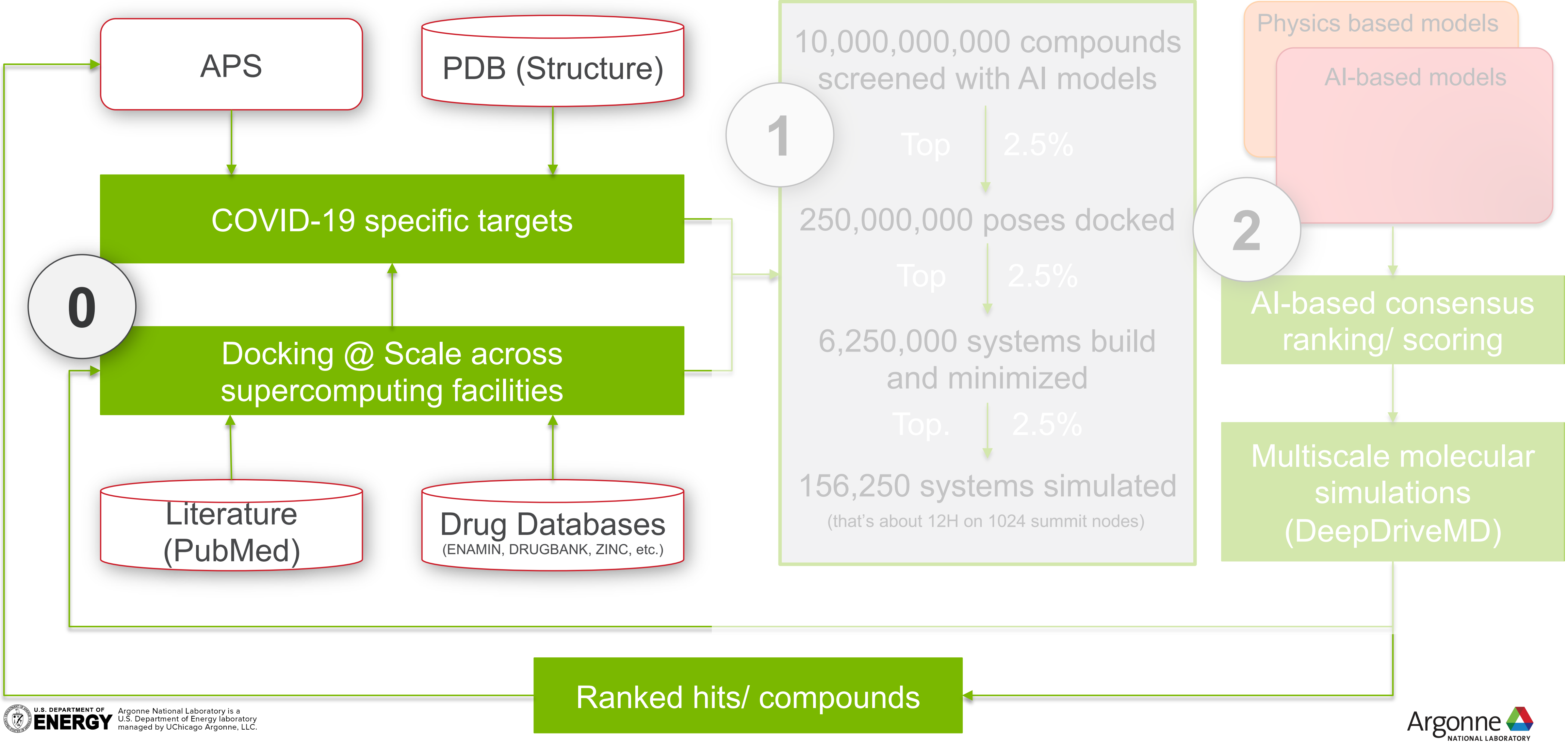
INTRODUCTION TO COVID-19 AND SARS-COV-2

- Observed first in Wuhan (Dec 2019)
  - Quickly spread to the province of Hubei and then onto the world
- Spreads via close contact or through respiratory particles
- Virus is larger and far more stable than its counterparts (SARS and MERS)
  - can live on surfaces for a while
- Need a comprehensive strategy to identify small molecules (or other therapeutic strategies) to treat infection



Veronica Falconieri Hays; Source: Lorenzo Casalino, Zied Gaieb and Rommie Amaro, U.C. San Diego (*spike model with glycosylations*) https://www.scientificamerican.com/article/a-visual-guide-to-the-sars-cov-2-coronavirus/

USING AI/ML TO DISCOVER DRUGS THAT CAN TARGET SARS-COV-2 PROTEOME



0

APS

PDB (Structure)

COVID-19 specific targets

Docking @ Scale across supercomputing facilities

Literature (PubMed)

Drug Databases
(ENAMIN, DRUGBANK, ZINC, etc.)

1

10,000,000,000 compounds screened with AI models

Top 2.5%

250,000,000 poses docked

Top 2.5%

6,250,000 systems build and minimized

Top. 2.5%

156,250 systems simulated
(that's about 12H on 1024 summit nodes)

2

Physics based models

AI-based models

AI-based consensus ranking/ scoring

Multiscale molecular simulations (DeepDriveMD)

Ranked hits/ compounds

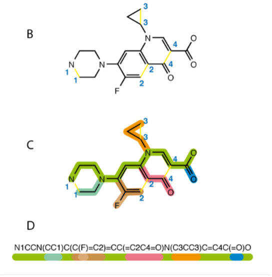
FIRST RELEASE OF HPC-COMPUTED FEATURES FOR AI-BASED DRUG SCREENING

23 input datasets, **4.2B molecules**, **60 TB** of molecular features and representations

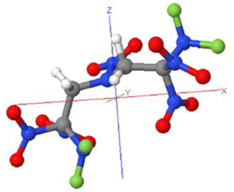
Data processing pipeline used ~2M core hours on ALCF Theta, TACC Frontera, OLCF Summit

1. Convert each molecule to a **canonical SMILES**
2. For each molecule, compute:
    a. ~1800 2D and 3D **molecular descriptors** using Mordred
    b. **Molecular fingerprints** encoding structure
    c. **2D images** of the molecular structure

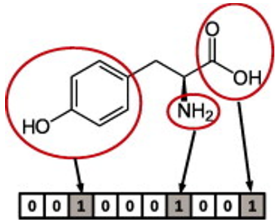
Computed data provide **crucial input features to AI models** for predicting molecular properties such as docking scores and toxicity



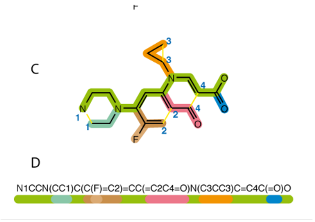
Canonical SMILES
23 CSV files with 4.2B molecules



Mordred Descriptors
420,130 CSV files, 48.70TB



Molecular Fingerprints
4,221 CSV files with base64 encoded fingerprints, 578.27GB



2D images
420,707 Pickle GZ files, 11.48 TB

https://2019-ncovgroup.github.io/data/

THE COVID'19 DATA PIPELINE:

USING AI AND SUPERCOMPUTERS TO ACCELERATE DRUG DEVELOPMENT

CHEMICAL LIBRARY DATABASE

4B known molecules

Enamine

DRUGBANK

BindingDB

GDB

eMolecules

cureFFI

MOSES

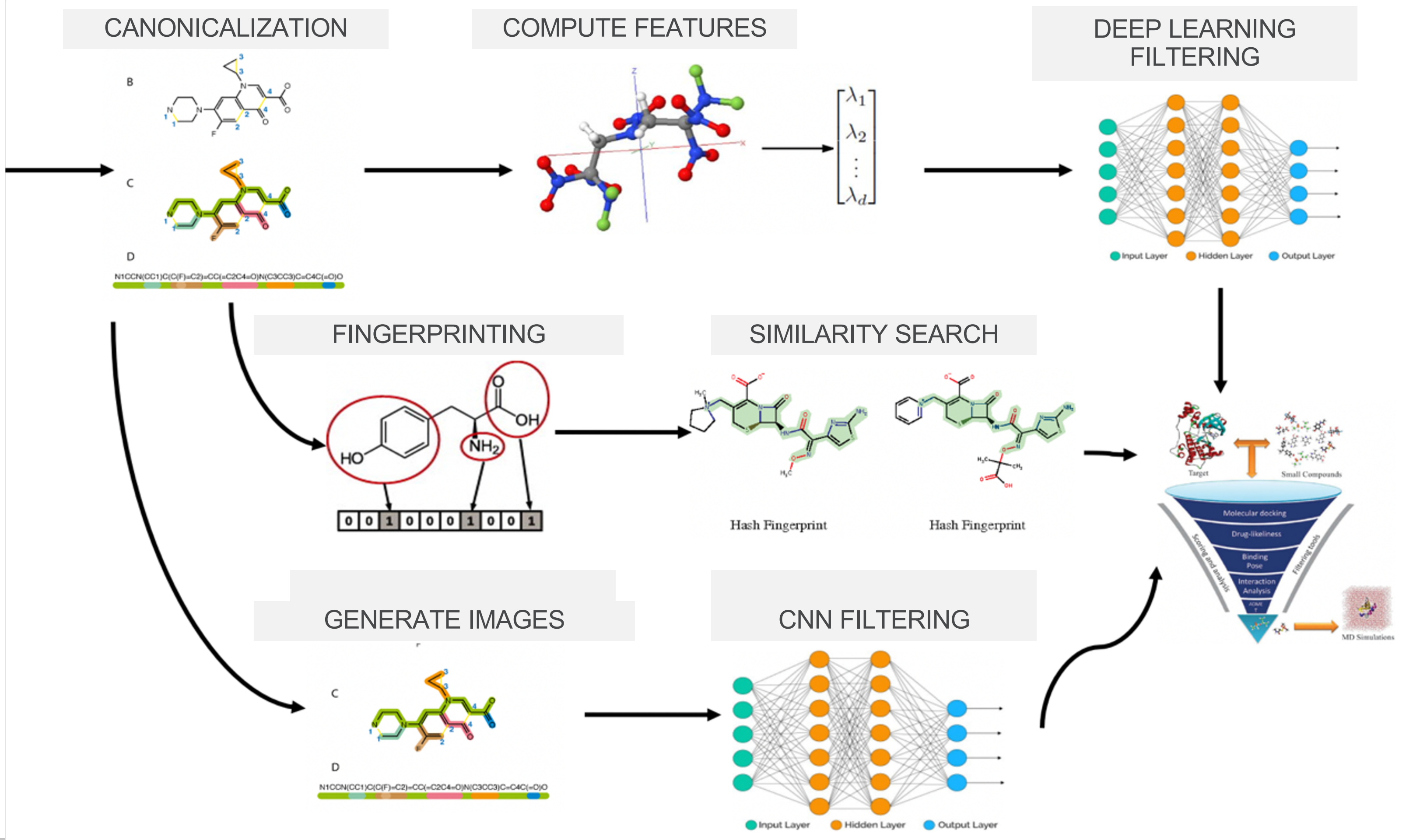
ZINC15

LINCS

SureChEMBL

PubChem

AND MORE



CANONICALIZATION

COMPUTE FEATURES

DEEP LEARNING FILTERING

FINGERPRINTING

SIMILARITY SEARCH

GENERATE IMAGES

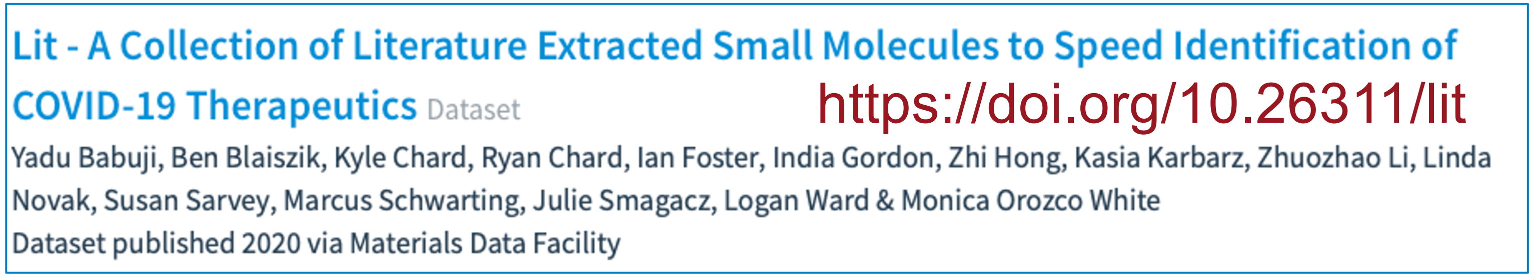
CNN FILTERING

COMPUTING RESOURCES

NATURAL LANGUAGE PROCESSING: DATASET AND CODE

Manual Extraction:

- Engaged Argonne CELS admin staff to extract small molecules from key SARS/SARS-CoV-2/MERS papers
- Extracted >800 molecules, structures



Lit - A Collection of Literature Extracted Small Molecules to Speed Identification of COVID-19 Therapeutics Dataset

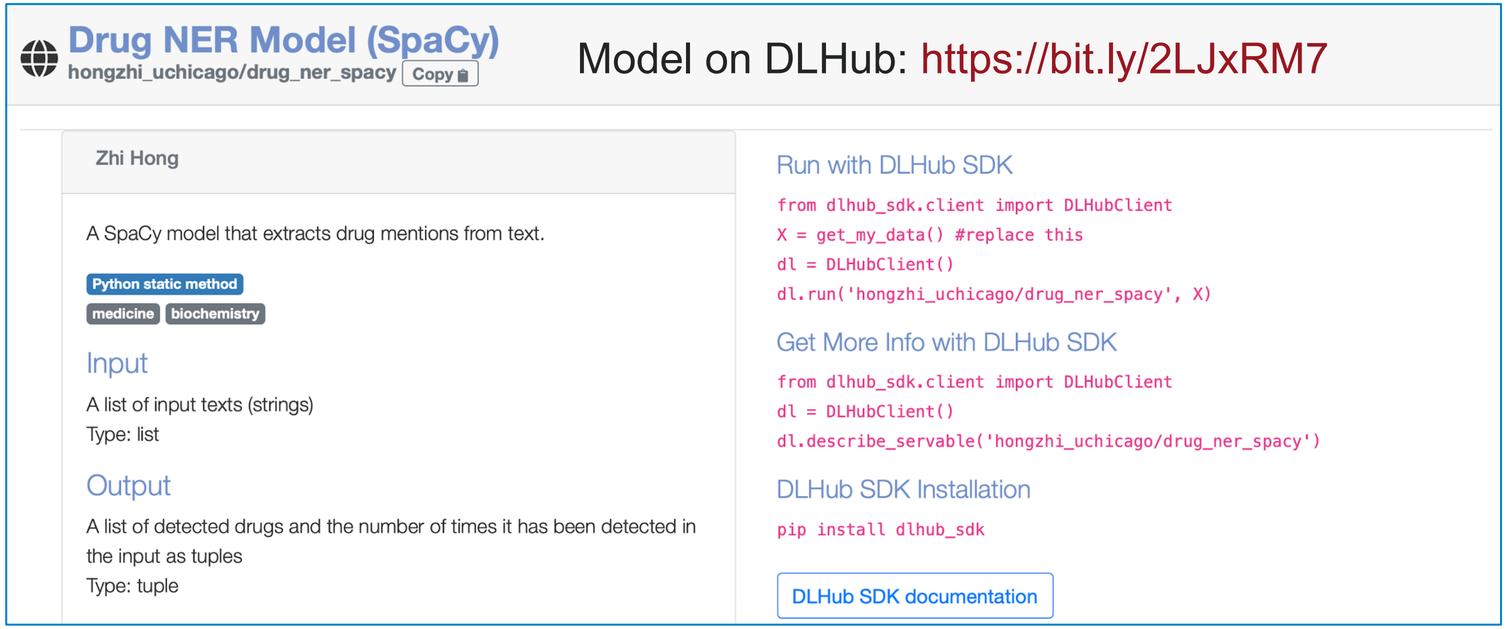
https://doi.org/10.26311/lit

Yadu Babuji, Ben Blaiszik, Kyle Chard, Ryan Chard, Ian Foster, India Gordon, Zhi Hong, Kasia Karbarz, Zhuozhao Li, Linda Novak, Susan Sarvey, Marcus Schwarting, Julie Smagacz, Logan Ward & Monica Orozco White

Dataset published 2020 via Materials Data Facility

Automated Extraction:

- Labeled relevant small molecules in their natural language context in CORD-19 papers
- Built named deep-learning entity recognition (NER) models to extract drug references from entire corpus (>24k full text articles)



Model on DLHub: https://bit.ly/2LJxRM7

Zhi Hong

A SpaCy model that extracts drug mentions from text.

Input

A list of input texts (strings)
Type: list

Output

A list of detected drugs and the number of times it has been detected in the input as tuples
Type: tuple

Run with DLHub SDK

```
from dlhub_sdk.client import DLHubClient
X = get_my_data() #replace this
dl = DLHubClient()
dl.run('hongzhi_uchicago/drug_ner_spacy', X)
```

Get More Info with DLHub SDK

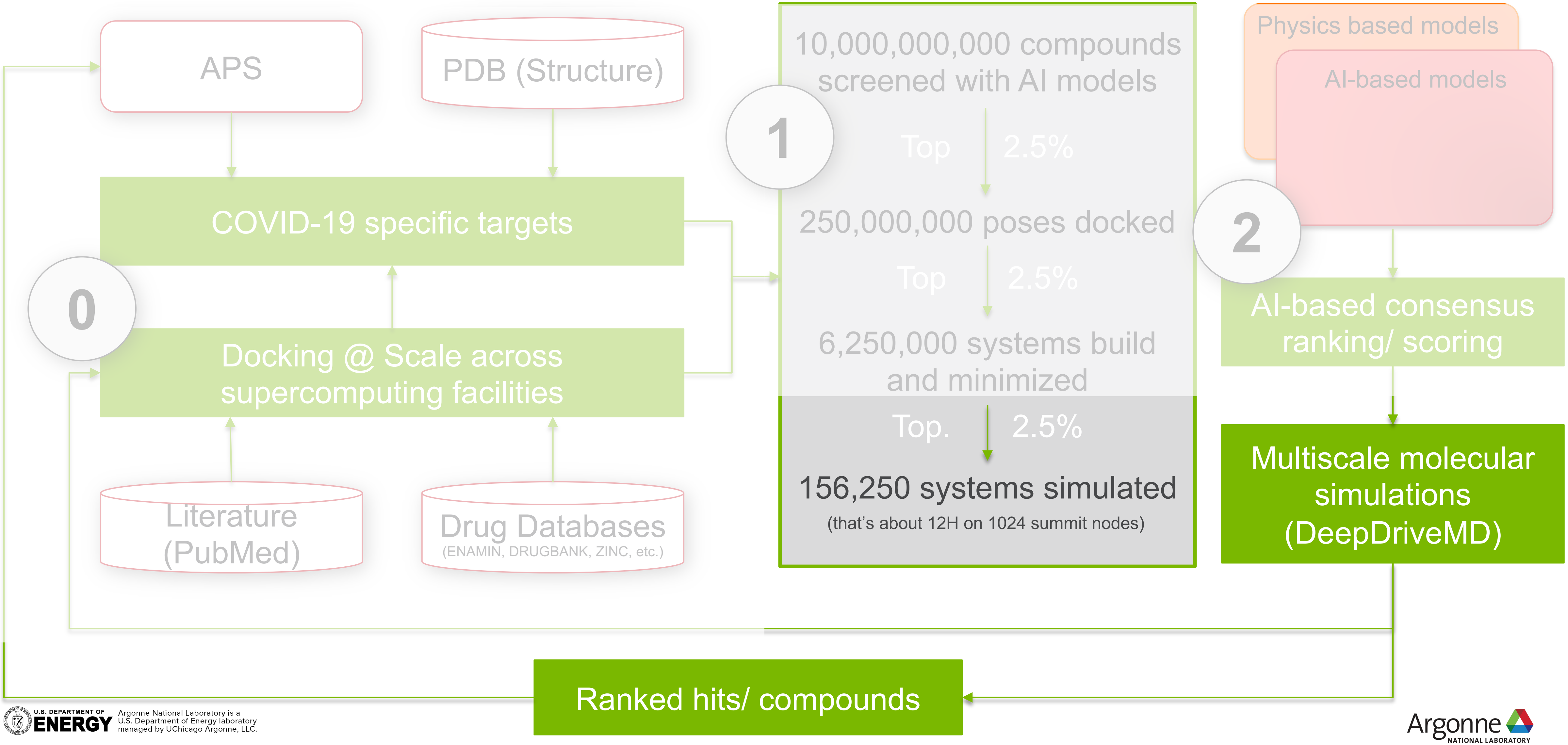
```
from dlhub_sdk.client import DLHubClient
dl = DLHubClient()
dl.describe_servable('hongzhi_uchicago/drug_ner_spacy')
```

DLHub SDK Installation

```
pip install dlhub_sdk
```

Code, training data: https://github.com/globus-labs/covid-nlp

USING AI/ML TO DISCOVER DRUGS THAT CAN TARGET SARS-COV-2 PROTEOME



0

APS

PDB (Structure)

COVID-19 specific targets

Docking @ Scale across supercomputing facilities

Literature (PubMed)

Drug Databases
(ENAMIN, DRUGBANK, ZINC, etc.)

1

10,000,000,000 compounds screened with AI models

Top 2.5%

250,000,000 poses docked

Top 2.5%

6,250,000 systems build and minimized

Top. 2.5%

156,250 systems simulated
(that's about 12H on 1024 summit nodes)

2

Physics based models

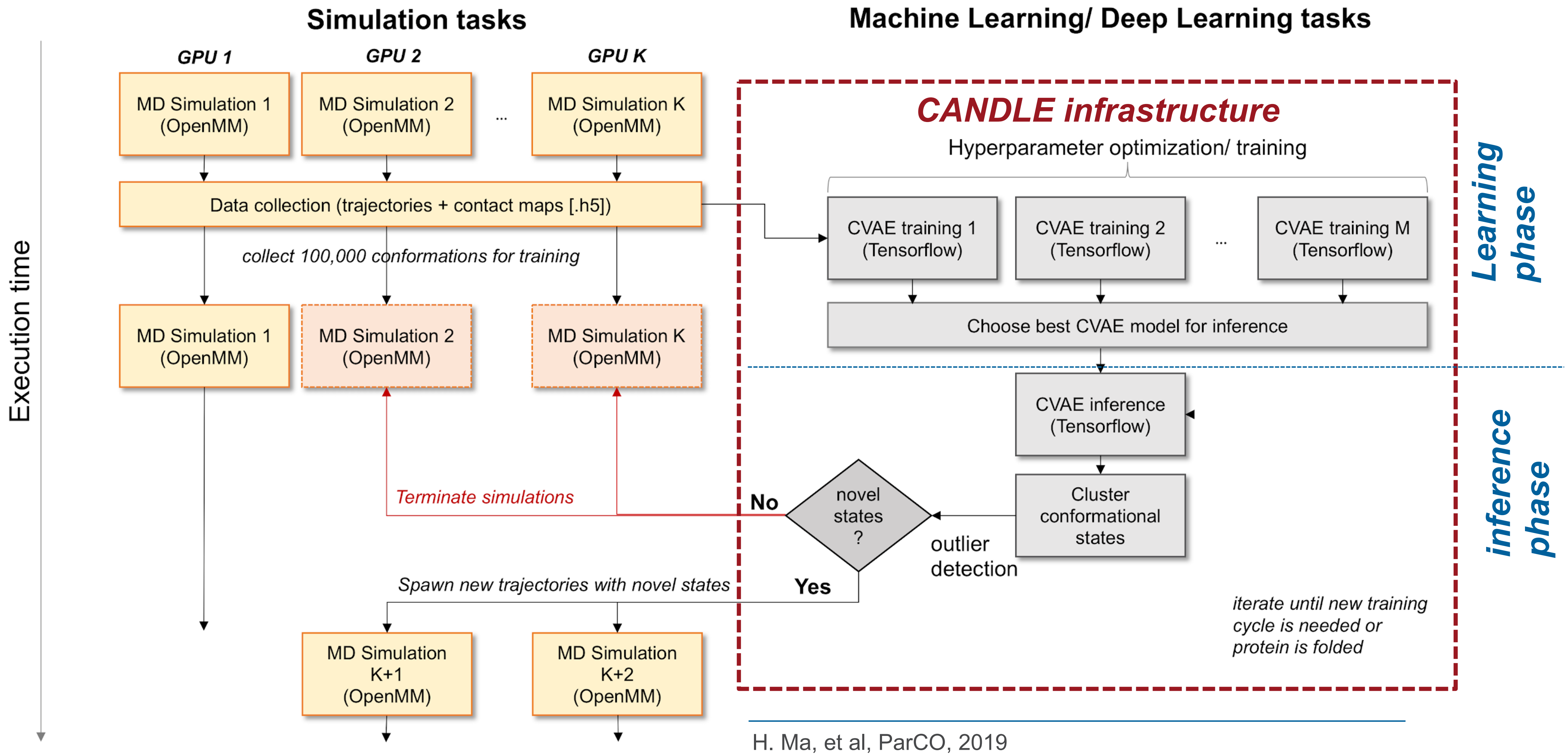
AI-based models

AI-based consensus ranking/ scoring

Multiscale molecular simulations (DeepDriveMD)

Ranked hits/ compounds

DEEPDRIVEMD: DL DRIVEN ADAPTIVE ENSEMBLES MD


Simulation tasks
Machine Learning/ Deep Learning tasks
Execution time
GPU 1
GPU 2
GPU K
MD Simulation 1 (OpenMM)
MD Simulation 2 (OpenMM)
...
MD Simulation K (OpenMM)
Data collection (trajectories + contact maps [.h5])
collect 100,000 conformations for training
MD Simulation 1 (OpenMM)
MD Simulation 2 (OpenMM)
MD Simulation K (OpenMM)
Terminate simulations
No
Spawn new trajectories with novel states
Yes
MD Simulation K+1 (OpenMM)
MD Simulation K+2 (OpenMM)
CANDLE infrastructure
Hyperparameter optimization/ training
CVAE training 1 (Tensorflow)
CVAE training 2 (Tensorflow)
...
CVAE training M (Tensorflow)
Choose best CVAE model for inference
Learning phase
CVAE inference (Tensorflow)
Cluster conformational states
novel states ?
outlier detection
inference phase
iterate until new training cycle is needed or protein is folded

H. Ma, et al, ParCO, 2019
H. Ma, et al, Workshop on Deep Learning on Supercomputers, 2019

Collaboration with Shantenu Jha (Rutgers/ Brookhaven) and RADICAL team

DEEPDRIVEMD OVERVIEW: INTERLEAVE SIMULATIONS AND ANALYTICS ADAPTIVELY FOR REDUCING COMPUTING OVERHEADS

Tr ons

“Big ir

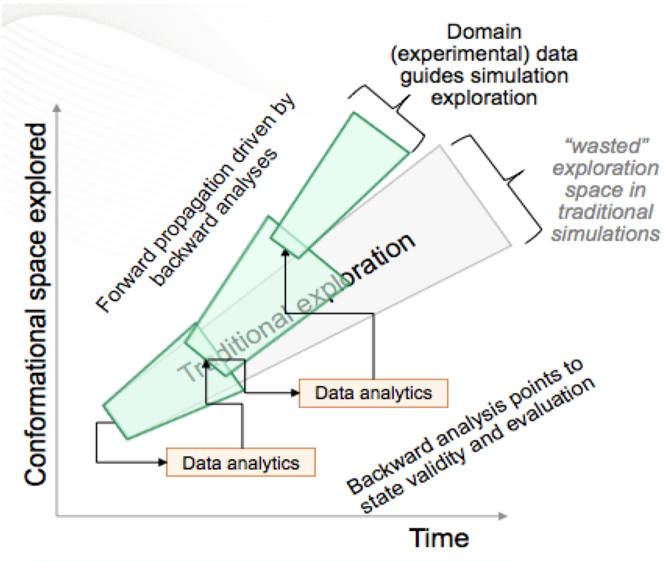
Big Store

Dedicate analytic clusters

ytics ata and eads itoring ck

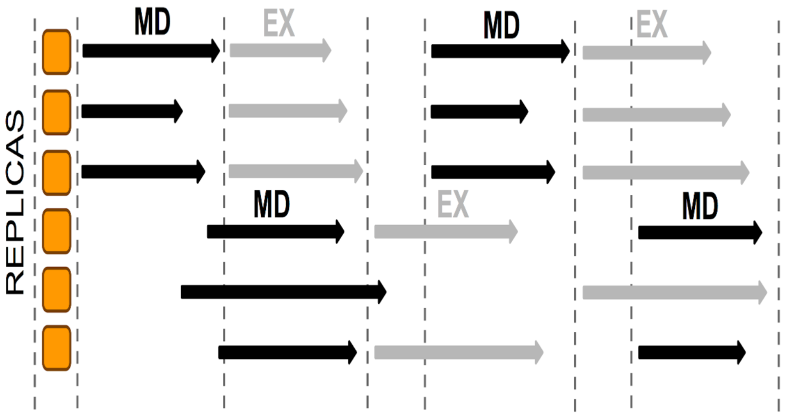
ze cation to ramework

- Generate ensemble of simulations in parallel as opposed to one realization of process
  - Statistical approach: **O(10^6 - 10^8)**!
- Ensemble methods necessary, not sufficient!
  - Adaptive Ensembles: **Intermediate data, determines next stages**
- Adaptivity: How, What
  - Internal data: Simulation generated data used to determine “optimal” adaptation

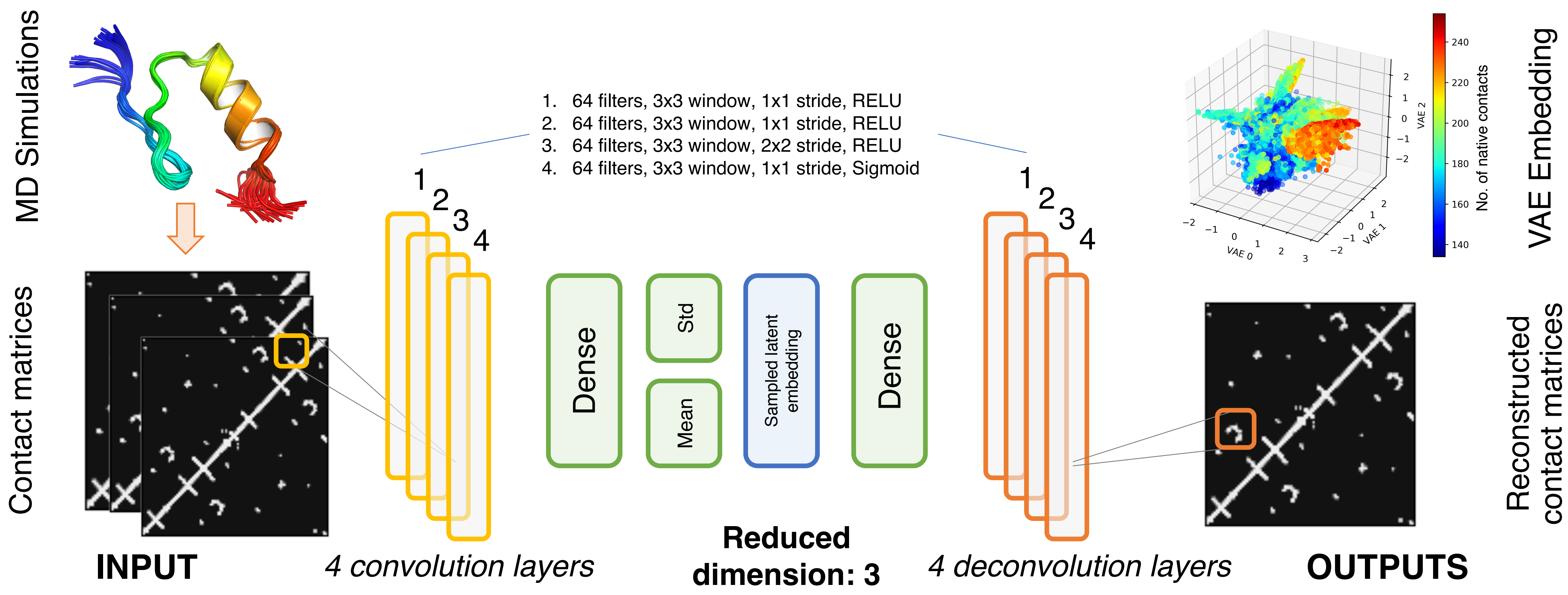


Domain (experimental) data guides simulation exploration
"wasted" exploration space in traditional simulations
Forward propagation driven by backward analyses
Traditional exploration
Conformational space explored
Data analytics
Data analytics
Backward analysis points to state validity and evaluation
Time

*Chodera, J.D., Noe, F., Curr. Opin. Struct. Biol. (2014)



A VARIATIONAL APPROACH TO ENCODE PROTEIN FOLDING WITH CONVOLUTIONAL AUTO-ENCODERS



MD Simulations

Contact matrices

INPUT

1. 64 filters, 3x3 window, 1x1 stride, RELU
2. 64 filters, 3x3 window, 1x1 stride, RELU
3. 64 filters, 3x3 window, 2x2 stride, RELU
4. 64 filters, 3x3 window, 1x1 stride, Sigmoid

4 convolution layers

Dense

Std

Mean

Sampled latent embedding

Dense

Reduced dimension: 3

4 deconvolution layers

VAE Embedding

Reconstructed contact matrices

OUTPUTS

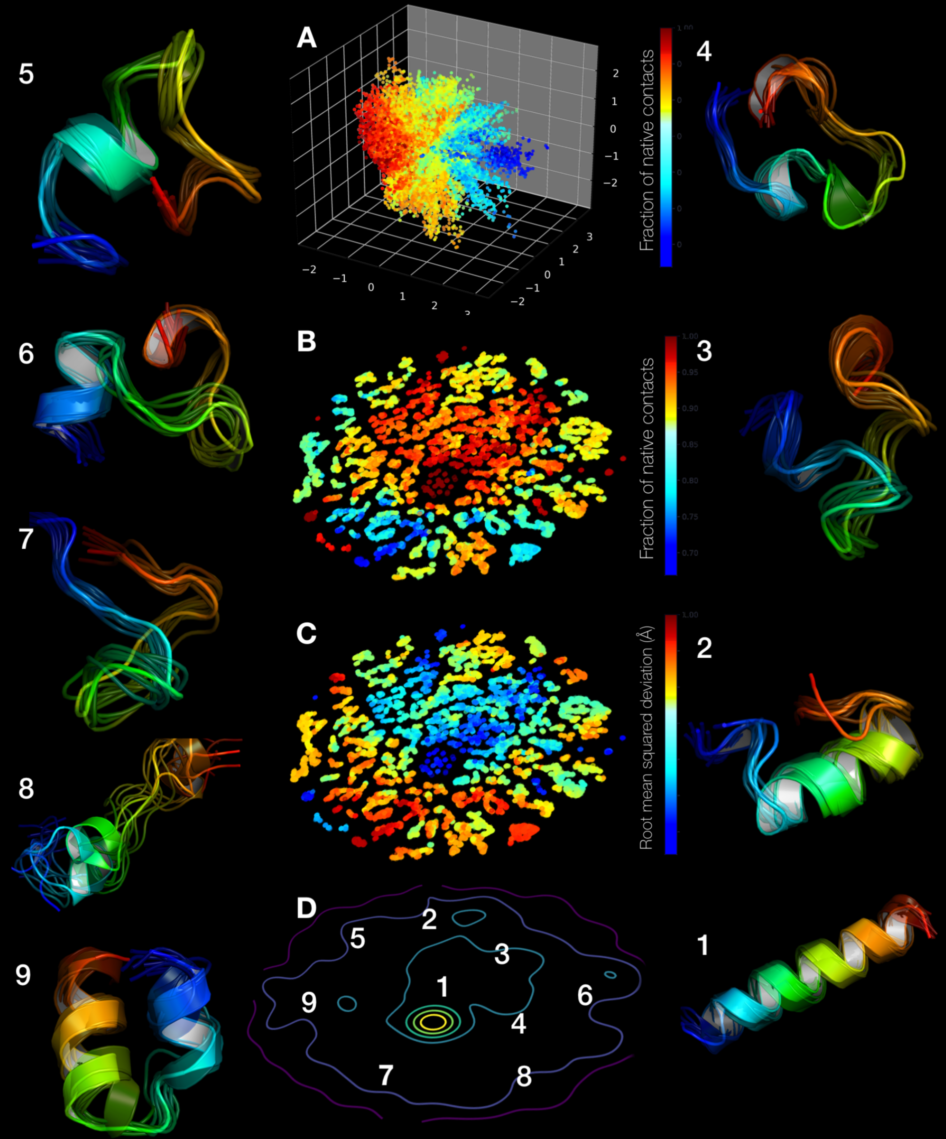
Related work:
Hernandez 17 arXiv,
Doerr 17 arXiv

DEEP CLUSTERING OF PROTEIN FOLDING SIMULATIONS

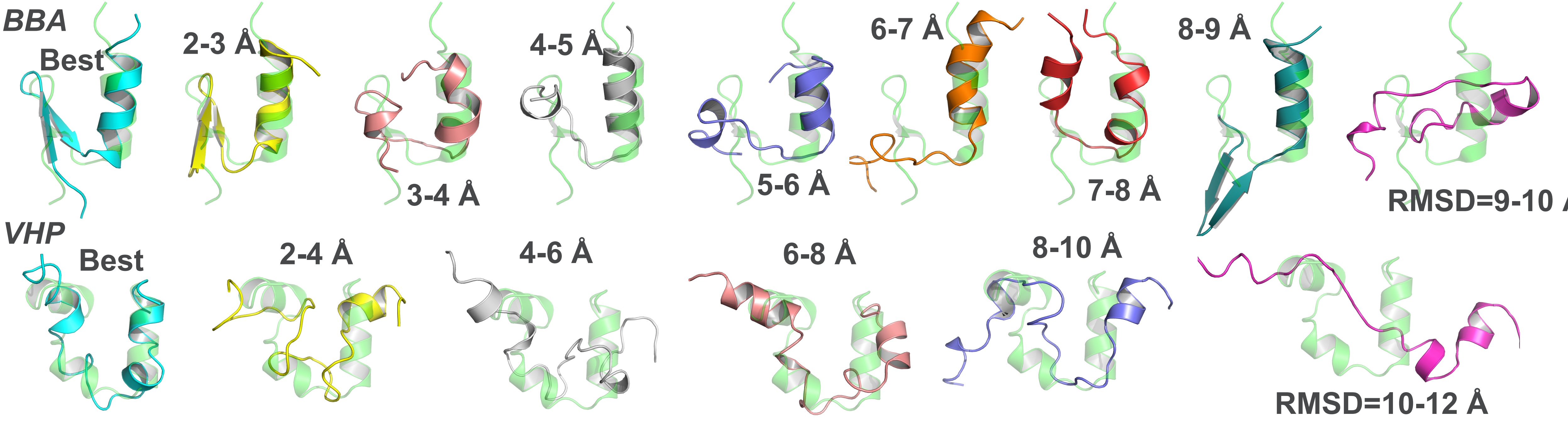
- Convolutional Variational Auto Encoders (CVAE)
  - Low dimensional representations of states from simulation trajectories.
  - CVAE can transfer learned features to reveal novel states across simulations

- Integrating Bayesian learning to support uncertainty in sampling novel states
  - HPC Challenge (1): DL approaches to achieve near real-time training & prediction!
  - HPC Challenge (2): Hyperparameter optimization (while model is training)!

Bhowmik, D., et al, BMC Bioinformatics (2018).

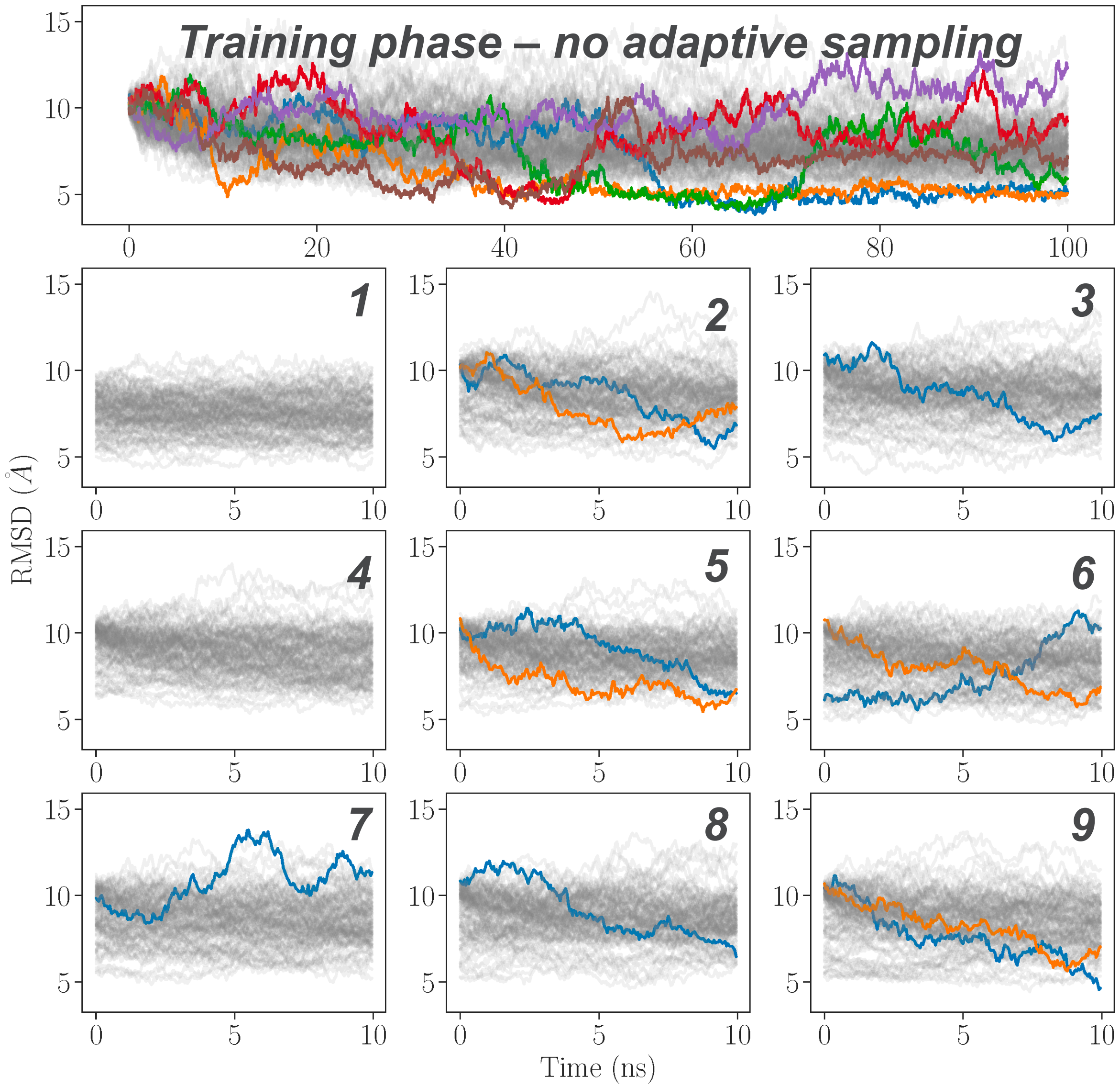


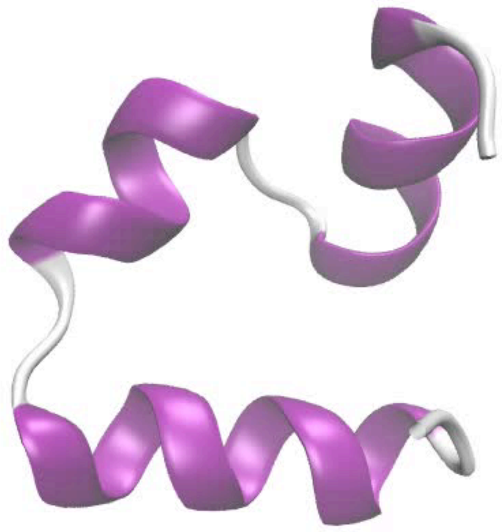
LARGER NUMBER OF SIMULATIONS IMPROVES FOLDING EFFECTIVENESS (HENCE SAMPLING)



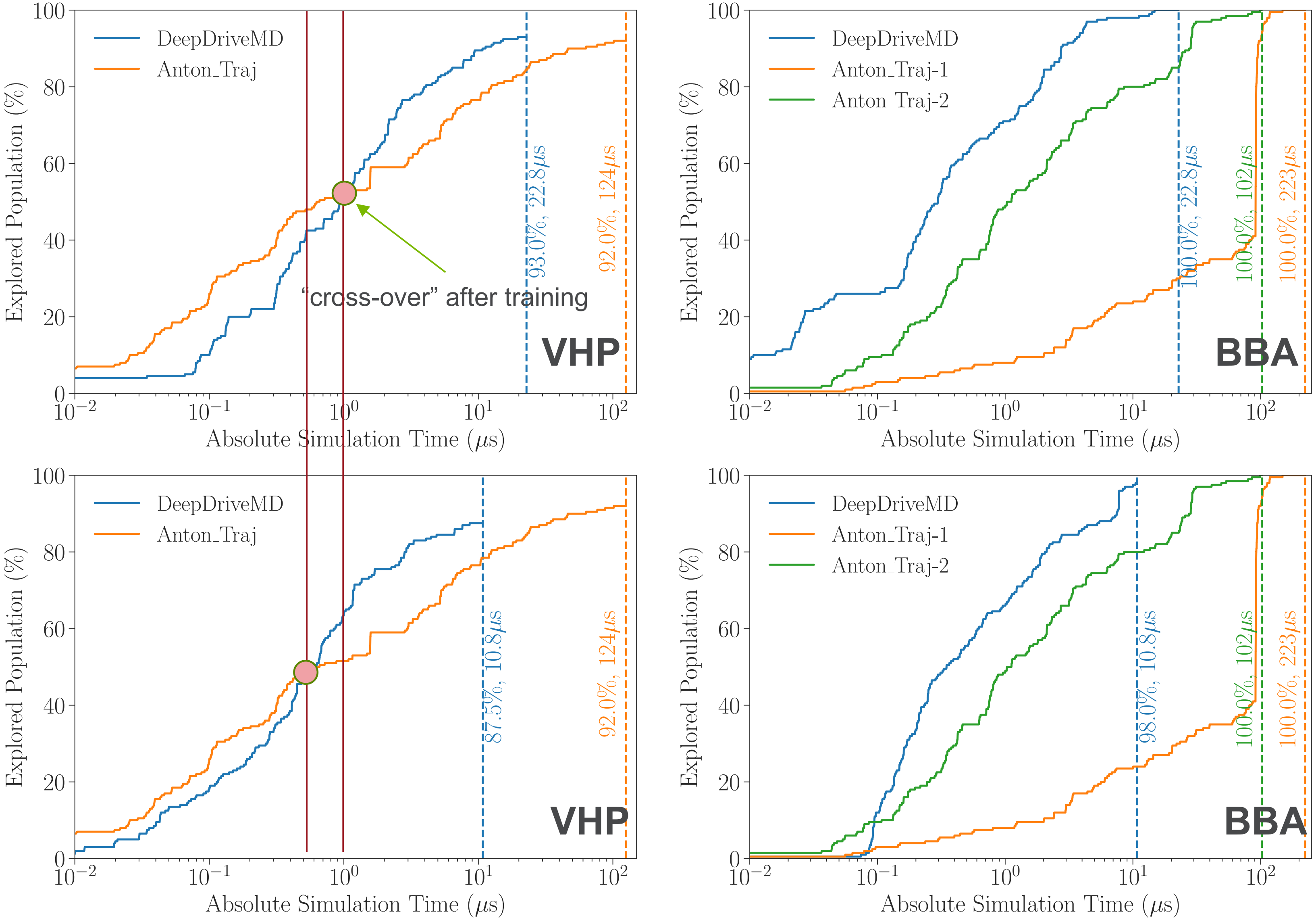
System	Total no. of simulations	Total simulation time (us)	First, subsequent simulations	Iterations	Min. RMSD
Fs-peptide	840	18.2	100, 10	7	0.29
BBA (FSD-EY)	1200	22.8	100, 10	10	1.8
VHP	1200	22.8	100, 10	10	3.83

ITERATIVE EXPLORATION OF STATES WITH DEEP LEARNING PROVIDES ACCESS TO FOLDED STATES





DEEPDRIVEMD SHOWS AT LEAST AN ORDER OF MAGNITUDE EFFICIENT SAMPLING COMPARED TO TRADITIONAL APPROACHES



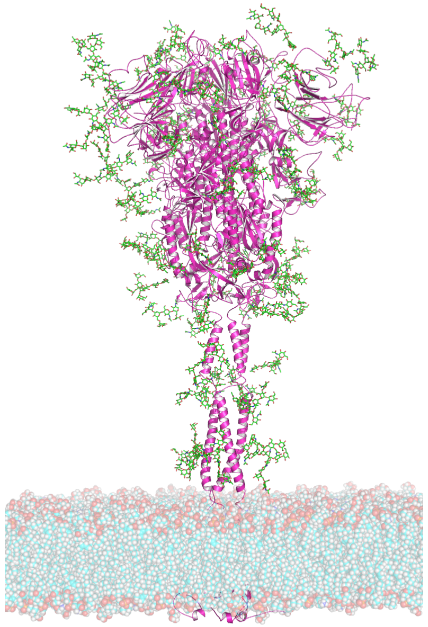
- **including the data from the "learning phase"**: one order of magnitude improvement in sampling:
  - Distinct "cross-over" after training where sampling is accelerated significantly after learning/ estimating the conformational states

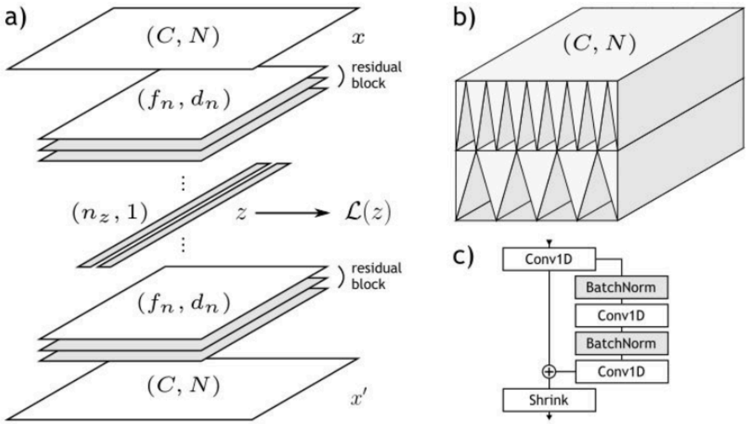
Reference trajectories are from D.E. Shaw (Science, 2011)

- ***not including the data from "learning phase":*** At least two orders of magnitude improvement in sampling:
  - If Anton trajectories take O(microsecond) to sample a particular state, DeepDriveMD samples it in O(100 ns)
- For BBA, 98% sampled states are observed within 10 microseconds!

Argonne
NATIONAL LABORATORY

USING FULLY CONVOLUTIONAL VAE TO IDENTIFY CONFORMATIONAL STATES IN SPIKE PROTEIN SIMULATIONS



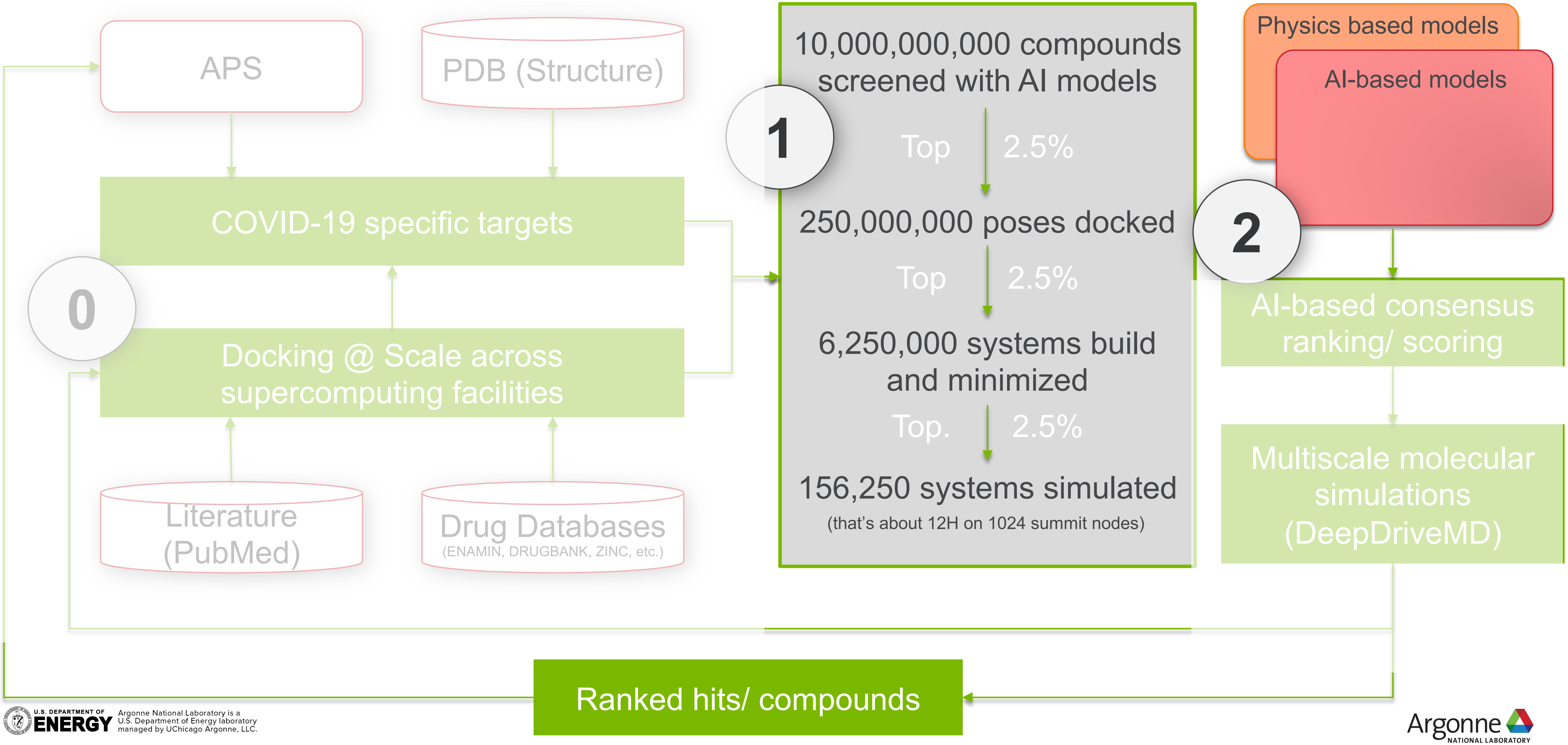


Total no. of parameters – 1.14 billion!

No. GPUs (V100)	Memory	Time per batch (8)
1	20213/32510 MiB	7.561
2	9947/32510 MiB (Encoder) 12987/32510 MiB (Decoder)	7.481

- Modification of the VAE architecture to accommodate larger systems (E.g. Spike protein – 1.5 million atoms)
- Model parallel example:
  - encoder and decoder on individual GPUs
  - implemented with Pytorch
- Can improve performance with layer-wise adaptive rescaling
- Joint work with Alex Brace (Argonne intern), Abe Stern (NVIDIA), Anda Trifan (CSGF), Rommie Amaro (UCSD), Carlos Simmerling (Stony Brook University)

USING AI/ML TO DISCOVER DRUGS THAT CAN TARGET SARS-COV-2 PROTEOME



0

APS

PDB (Structure)

COVID-19 specific targets

Docking @ Scale across supercomputing facilities

Literature (PubMed)

Drug Databases
(ENAMIN, DRUGBANK, ZINC, etc.)

1

10,000,000,000 compounds screened with AI models

Top 2.5%

250,000,000 poses docked

Top 2.5%

6,250,000 systems build and minimized

Top. 2.5%

156,250 systems simulated
(that's about 12H on 1024 summit nodes)

2

Physics based models

AI-based models

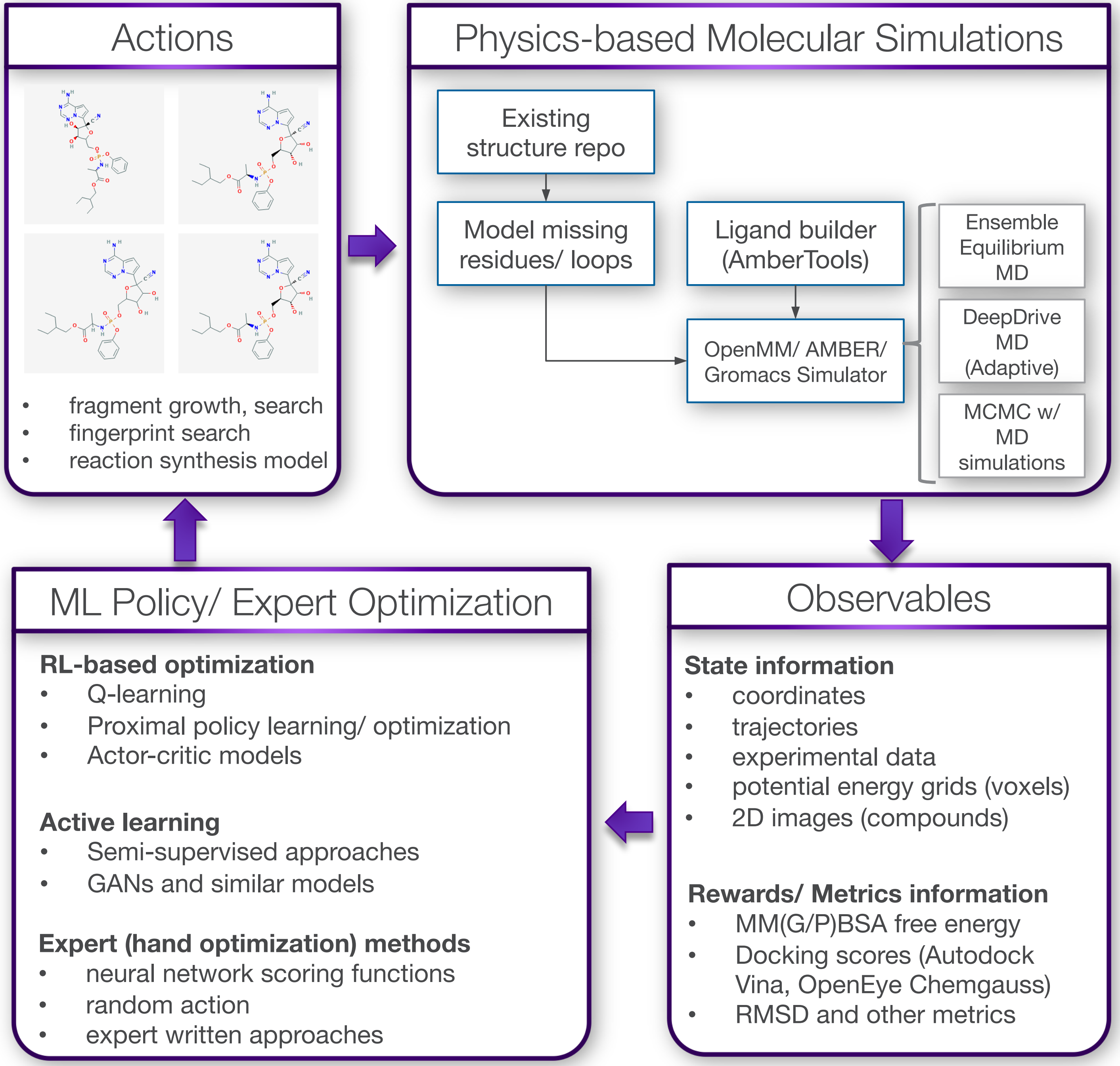
AI-based consensus ranking/ scoring

Multiscale molecular simulations (DeepDriveMD)

Ranked hits/ compounds

REINFORCEMENT LEARNING DRIVEN MD

- Motivation: physics-based models are guided by an action space determined by AI
- Can we expand the compound space explored using RL?
- For SARS-CoV-2 proteome:
  - relevant for specific mutations compared to other CoV proteins
  - suggest repurposing based on shape/structural complementarity



Actions

- fragment growth, search
- fingerprint search
- reaction synthesis model

Physics-based Molecular Simulations

Existing structure repo

Model missing residues/ loops

Ligand builder (AmberTools)

OpenMM/ AMBER/ Gromacs Simulator

Ensemble Equilibrium MD

DeepDrive MD (Adaptive)

MCMC w/ MD simulations

Observables

State information

- coordinates
- trajectories
- experimental data
- potential energy grids (voxels)
- 2D images (compounds)

Rewards/ Metrics information

- MM(G/P)BSA free energy
- Docking scores (Autodock Vina, OpenEye Chemgauss)
- RMSD and other metrics

ML Policy/ Expert Optimization

RL-based optimization

- Q-learning
- Proximal policy learning/ optimization
- Actor-critic models

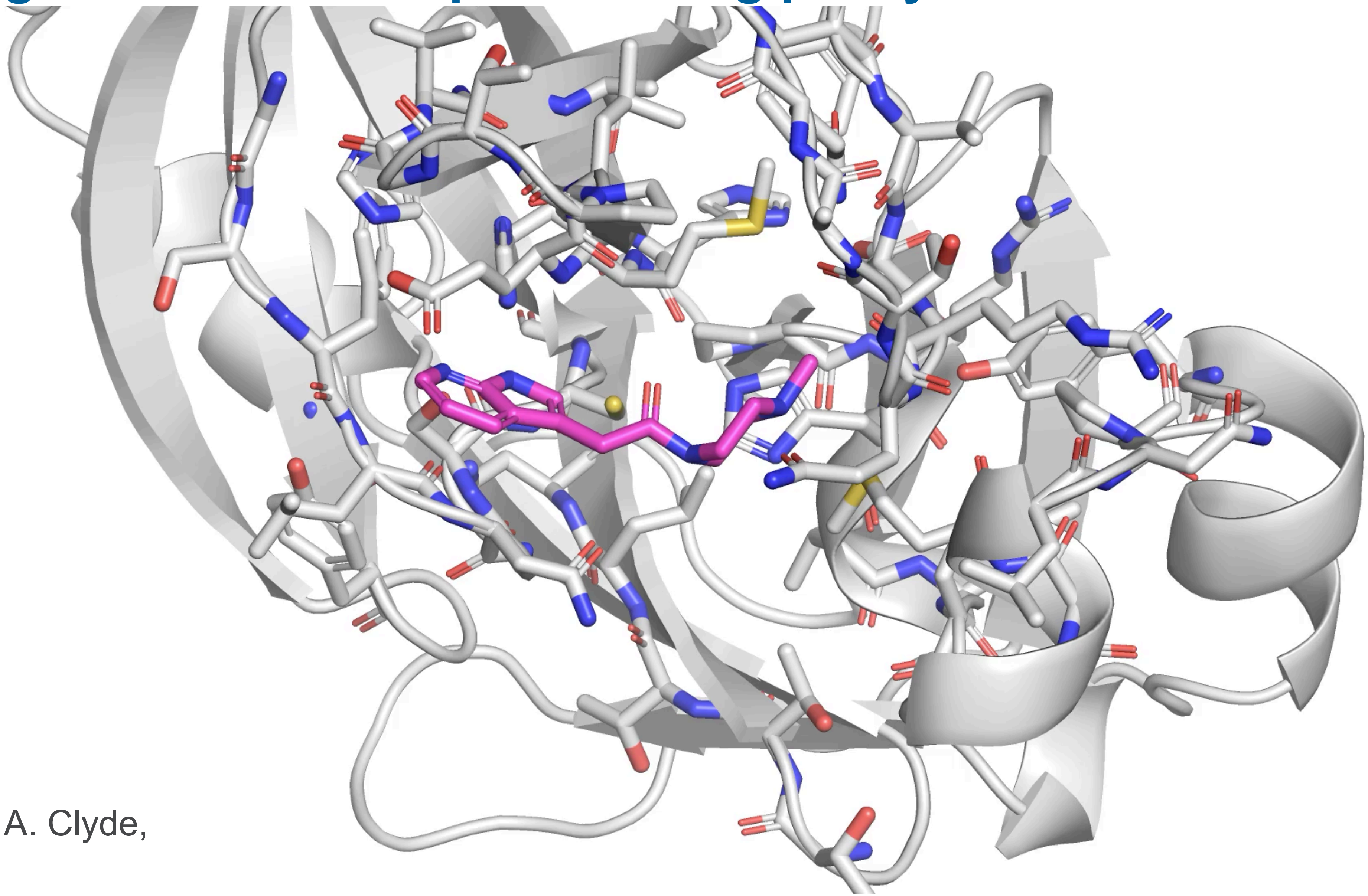
Active learning

- Semi-supervised approaches
- GANs and similar models

Expert (hand optimization) methods

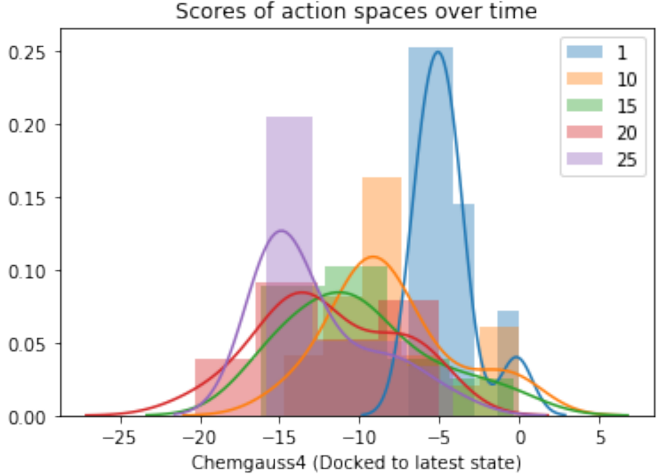
- neural network scoring functions
- random action
- expert written approaches

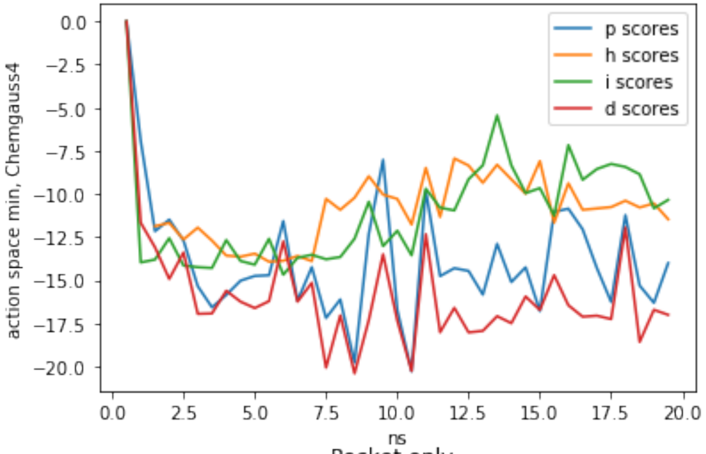
Fragment growth with an expert docking policy

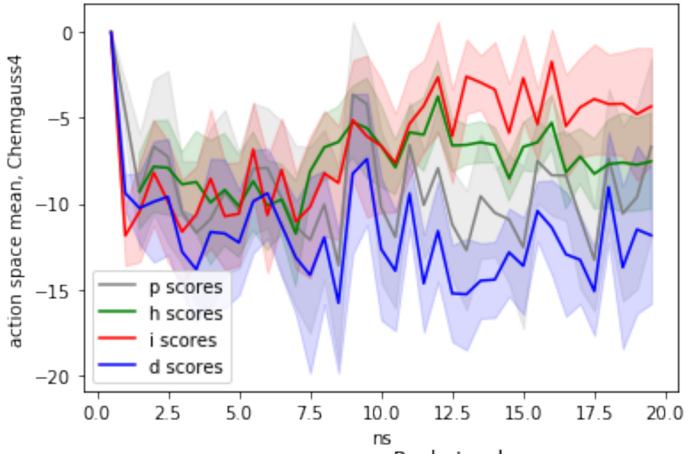


Joint work with A. Clyde, UChicago

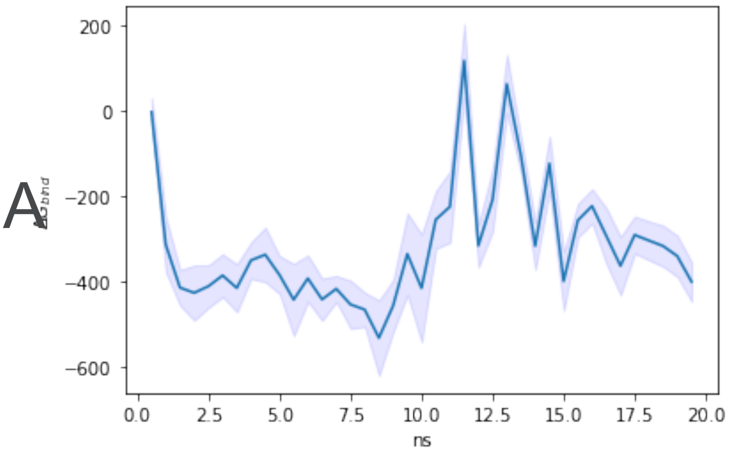
Action space

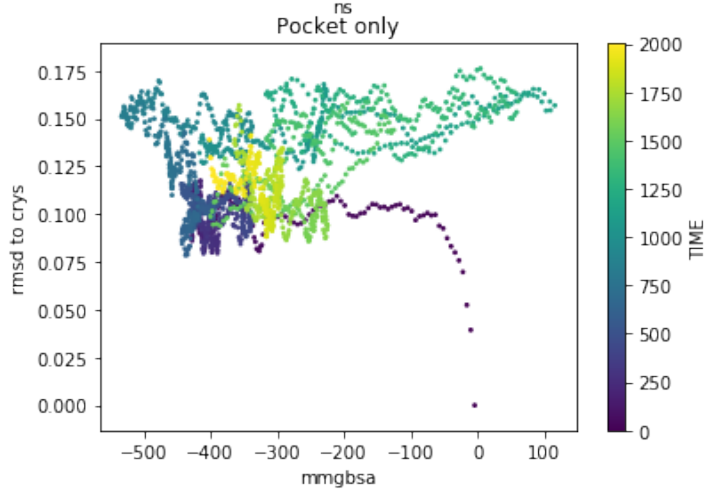


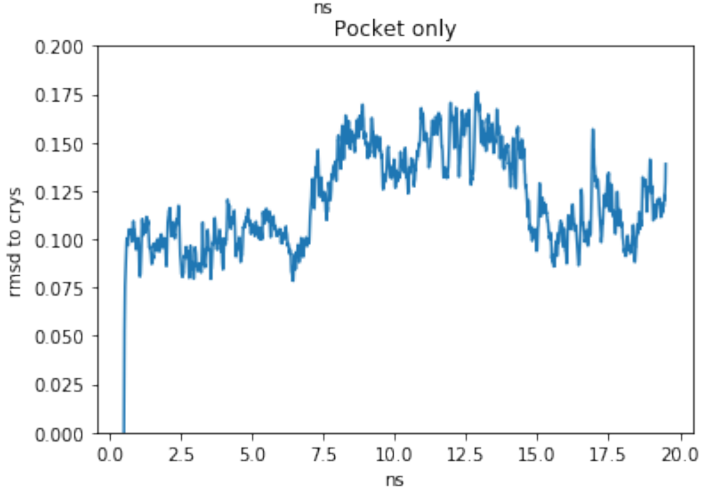




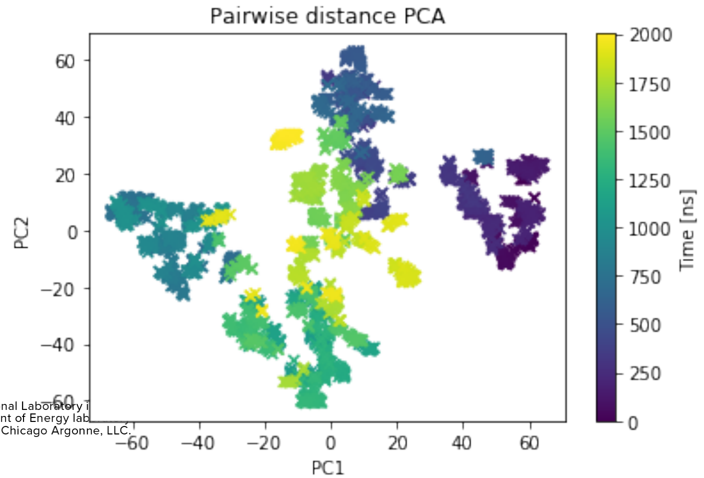
MMGBSA

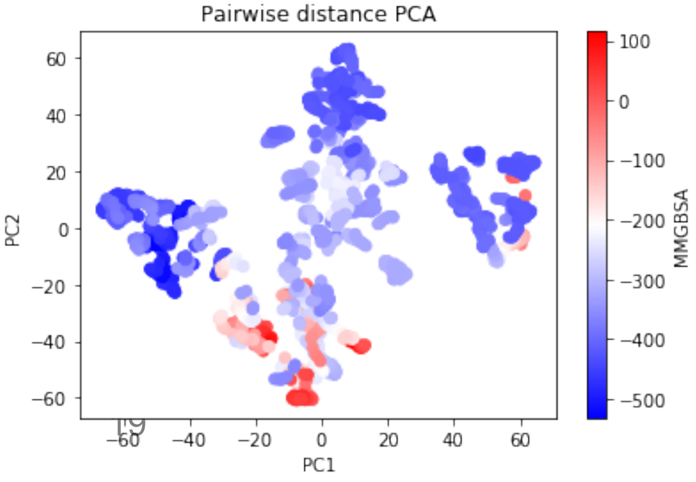


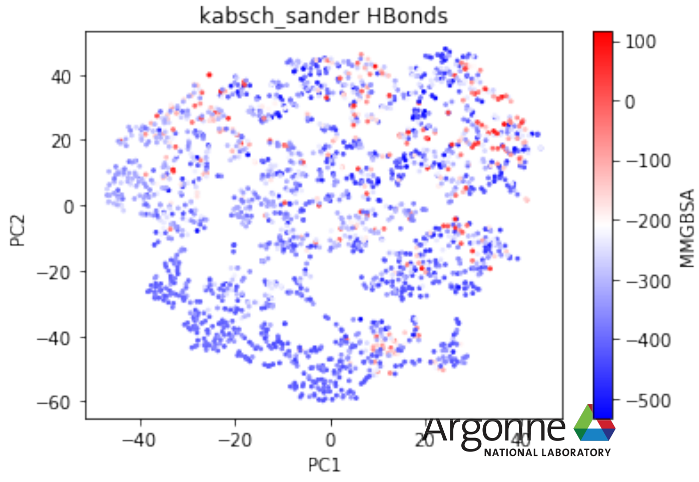




Sampling







FUTURE WORK / OUTLOOK

Conformational landscapes of proteins:

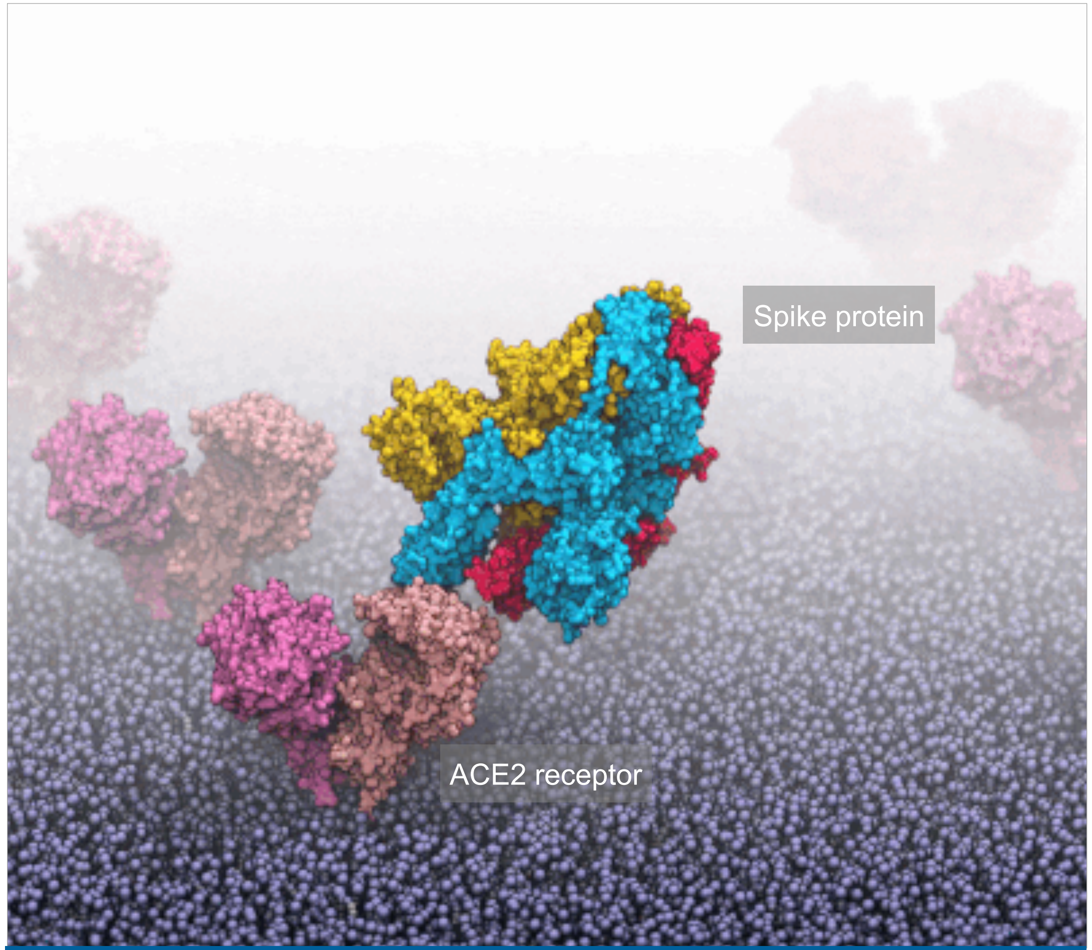
- *Sampling remains challenging*: are there techniques that can aid accurate biophysical characterization of protein conformational landscapes?
- *Deep learning / AI techniques show promise*: are they learning biophysical characteristics that can be used to guide simulations?
- *Protein interactions need "context"*: are there multi-scale methods to integrate information across experiments, simulations and theory?

AI/ML coupled to simulations (challenges)

- Improvement in additional AI/ML models
- Active learning approaches for docking ligands
- Runtime systems are unprepared for such use cases where AI/ML systems drive simulations :
  - improving exchange of data with concurrently running models
  - tracking datasets as simulations are running (online/ in situ training)

FUNDING AND ACKNOWLEDGEMENTS

- Everyone in the team (all ~300)
- Computing support:
  - ALCF, OLCF
  - TACC, SDSC, IU
  - HPC Consortium
- Funding acknowledgement:
  - DOE National Virtual Biotechnology Laboratory (NVBL)
  - Argonne internal funding (LDRD)
  - DOE Exascale computing project (Cancer Deep Learning Environment)



Simulations driven by AI depict how the CoV-2 spike protein attaches to the human ACE2 receptor protein

THANK YOU!

(RAMANATHANA@ANL.GOV)